Programcsomag információkinyerési kutatások támogatására

Alexin Zoltán[1], Gyimóthy Tibor[1], Csirik János[1]

Szegedi Tudományegyetem, TTK, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail:{alexin,gyimothy,csirik}@inf.u-szeged.hu

Kivonat A publikációban bemutatásra kerül egy információkinyerési kutatásokat támogató programcsomag, amelynek moduljai a nyers szöveg beolvasásától kezdve a végeredmény webes megjelenítéséig minden szükséges funkciót megvalósítanak. A modulok egymással szabványos TEI XML állományok segítségével kommunikálnak, amelyek a feldolgozás tetszőleges szakaszában elemezhetők. A technológia ezen a módon támogatást nyújt az egyes modulok önálló fejlesztéséhez és teszteléséhez. A fontosabb modulok: a szegmentáló, morfológiai elemző, szófaji egyértelműsítő, felszíni szintaktikai elemző, szemantikai bővítménykezelő, eseménymintákat felismerő mintaillesztő és webes megjelenítő modul. A szerzők a programcsomag működését egy kísérleti rendszeren mutatják be, amely üzleti rövidhírekből gyűjt különböző információkat. [1]

Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, felszíni szintaktikai elemzés

1. Bevezetés

Az információkinyerés (IE, Information Extraction) technológiájának kutatása dinamikusan fejlődő terület a természetesnyelv-feldolgozásban. Az Interneten megjelenő hatalmas információtömeg gépi feldolgozása és a kívánt információ tömör formában történő összegyűjtése napi szükséglet, amelyre a gazdaság, a tudomány, a politika, de akár a hírszerzés területén is van igény. Míg az információ visszakeresés (IR, Information Retrieval), amely a webes kereső programok jellemző tevékenysége, arra irányul, hogy a felhasználó igényeinek megfelelő dokumentumokat változatlan formában bocsássa rendelkezésre, addig az információkinyerés célja a megtalált dokumentumokban a lényeges információ megjelölése, majd összegyűjtése. A számítógéppel támogatott szövegtömörítés, kivonatolás és az információkinyerés szoros kapcsolatban áll egymással.

Az információkinyeréssel foglalkozó rendszerek nem törekednek a szövegek teljes megértésére és analízisére, a fő követelmény velük szemben a nagy kapacitás, a gyorsaság és egy elfogadható szintű pontosság. Rendszerint megelégednek

[1] A szerzők köszönetüket fejezik ki az Oktatási Minisztériumnak, amely az NKFP 2/017/2001 projekt keretében az itt ismertetésre kerülő kutatást támogatta.

a mondatok fontosabb szereplőinek azonosításával anélkül, hogy részletes szintaktikai elemzést végeznének. Ehhez egy ún. felszíni elemzés (shallow parsing) végrehajtására van szükség. A szereplők és mondatbeli szerepeik azonosításában fontos szerepet játszanak a tulajdonnevek (*named entityk*). A gyakori közéleti szereplők, cégek, európai és magyar városok nevének felismerése egy nagyméretű lexikon alapján, a szófaji elemzéstől függetlenül történik.

A továbbiakban bemutatásra kerül egy, a Szegedi Tudományegyetem Informatikai Tanszékcsoportjában kifejlesztett programcsomag, amely az információkinyerési kutatások támogatására készült. A koncepció egyik legfontosabb szempontja a modularitás volt, így a komponensek egymástól függetlenül fejleszthetők. A modulok egymástól függetlenül futtathatók, így a feldolgozás minden egyes lépése nyomon követhető és ellenőrizhető. A fenti tulajdonságok nagy súllyal esnek latba a kutatások kezdeti szakaszában, amikor az egyes modulok kísérleti fejlesztése folyik. Az egyes modulok szabványos kommunikációja megkönnyíti a modulok gyors cseréjét, a leghatékonyabb mód- szer kiválasztását.

Az elkészült rendszert gazdasági témájú rövidhírek feldolgozására alkalmazták. Az MTI-Eco, Business+ szolgáltatását [2] felhasználva egy 6453 hírből álló adatbázist hoztak létre, amely egyaránt szolgált tréning- és tesztelési célokat. A kutatási és fejlesztési munkában, az újabb, egyre tökéletesebb modulok készítése során egy hasonló rendszer nem nélkülözhető. A jelenlegi fejlesztések a kulcspozíciót betöltő modulokra irányulnak: a szóalaktani egyértelműsítő, a felszíni szintaktikai elemző és az eseménymintákat, ún. *szemantikus kereteket* felismerő modulra.

2. A programcsomag fontosabb tulajdonságai

Az ismertetésre kerülő programcsomag moduljai végigkövetik a nyers szöveg feldolgozásának egyes lépéseit a szöveg mondatokra és szavakra bontásától kezdve, a szóalaktani elemzésen, egyértelműsítésen, és a felszíni elemzésen át, a mondatminták felismeréséig, a szereplők azonosításáig, majd pedig az eredmények tömörített formában történő webes megjelenítéséig.

2.1. Az információkinyeréshez kapcsolódó modulok fejlesztését támogató adatbázis: a Szeged Korpusz 2.0

Az Oktatási Minisztérum által támogatott IKTA 27/2000 projekt keretében készült el a Szeged Korpusz 1.0-s változata[1], amely egy szófajilag elemzett, majd kézzel egyértelműsített adatbázis volt. Ezt az információkinyerési kutatások támogatására az MTA Nyelvtudományi Intézettel és a MorphoLogic Kft.-vel közös konzorcium jelentősen továbbfejlesztette. A Szeged Korpusz újabb változata [3] hat különböző témakörben gyűjtött, összesen 1,2 millió szót tartalmazó, számítógéppel feldolgozható szöveg. Ennek az állománynak egy mintegy 200 ezer

[2] Magyar Távírati Iroda, `http://www.mti.hu`

[3] SZTE, Informatikai Tanszékcsoport, Nyelvtechnológiai Csoport: `http://www.inf.u-szeged.hu/hlt`

szavas részét képezi a bevezetőben már említett 6453 MTI rövidhírt tartalmazó anyag.

A Szeged Korpusz 2.0[5][6] kialakítása során a készítők figyelembe vették egy majdani információkinyeréssel kapcsolatos alkalmazás fejlesztésekor felmerülő igényeket. Elsősorban a felszíni elemzés jelenségeinek tanulmányozása érdekében került sor a szövegben a főnévi szerkezetek teljes annotációjára, azaz minden egyes főnévi szerkezetet és annak belső is szerkezetét megjelölték. A kezdeti annotációt számítógépes program állította elő, amelyet azután nyelvész szakértők egyenként ellenőriztek, és hiba esetén javítottak. A szövegadatbázis nagy mennyiségben tartalmaz tulajdonneveket, a hírekben gyakran előforduló cégek és személyiségek neveit.

A korpusz rövidhíreit a fentieken kívül további információkkal is kiegészítették, nevezetesen az IPTC[4] által javasolt tematikus kódolással. A tematikus kódok alapján a gazdasági rövidhírek egy nagyon finom osztályozását kapták: csupán az üzleti élet témakörén belül 41 alkategóriát különböztettek meg például: könyvvizsgálás/auditálás, vállalati közgyűlés, éves beszámoló, adás-vétel stb. A 6453 rövidhírből 2478 hír kapcsolódott a vállalati üzleti élethez, a többi kereskedelmi, tőzsdei, pénzügyi és egyéb hír volt. A főnevekhez kapcsolódó lexikális ismeretek tárolására egy egyszerű felépítésű ontológiai adatbázist hoztak létre a konzorciumi partnerek. Ebből megtudható például, hogy egy adott főnév élőlény, élettelen tárgy, vagy pénznem-e, illetve tulajdonnévként személy, cég vagy bank neve-e. Az eseménymintákat felismerő modul nagy mértékben támaszkodik ezekre az információkra. Az ontológiai adatbázis jelentős fejlesztése folyamatban van.

2.2. A modulok szabványos XML felületen történő kommunikációja

A Szeged Korpusz 2.0 készítésekor a TEI [5] XML[6] dokumentumtárolási szabványt vették alapul. Ez az elektronikus szövegek tárolására szolgáló technológia széles körben elterjedt. Az Informatikai Tanszékcsoport a TEI-konzorcium alapító tagja, a TEI konzorcium honlapján 119 különböző projekt ismertetője található a világ minden tájáról, beleértve a Szeged Korpuszét is. Az XML-formátum lehetővé teszi, hogy a nyers szöveghez ún. metainformációkat rendeljenek, amelyeket címkék jeleznek a szövegben. Az 1. ábrán egy példa található erre: a `<sentence>` címke a mondatokat, a `<word>` a szavakat, míg az `<mscat>` címke a morfo-szintaktikai kategóriát jelzi. Általában a címkék opcionálisak, ezért ugyanahhoz a szöveghez a metainformációk egyre bővülő halmazát lehet hozzárendelni a feldolgozás előrehaladtával.

A természetes nyelvi feldolgozás, az ismertetett programcsomag moduljainak eredményei a nyers szövegbe metainformációként kerülnek be, például az alaktani elemzés eredménye, vagy a felszíni elemzés eredménye. Minden egyes modul növeli a kiindulási szöveg metainformáció-tartalmát. A TEI-szabvány a szükséges

[4] Az IPTC (International Press Telecommunication Council) által javasolt tematikus kódolás: `http://www.iptc.org/site/subject-codes/subjectcode.html`

[5] Text Encoding Initiative, `http://www.tei-c.org`

[6] XML információs oldal: `http://www.xml.org`

címkék leírásával, a metainformációk belső struktúrájának formalizálásával nyújt segítséget ehhez a feladathoz.

```
<xml>
  <sentence id="1.1">
    <word>A<mscat>NE</mscat></word>
    <word>kutya<mscat>FN</mscat></word>
    <word>ugat<mscat>IGE</mscat></word>
    <punctuation>.</punctuation>
  </sentence>
</xml>
```

1. ábra. Egy XML állomány részlete

3. A programcsomag fontosabb moduljai

A következőkben a programcsomag egyes moduljai kerülnek részletesebben is bemutatásra. A szerzők által vezetett kutatócsoport egyik fő célkitűzése az volt, hogy a modulok fejlesztéséhez lehetőség szerint gépi tanuló algoritmusokat alkalmazzanak. A tanuló rendszerek segítségével és a nyelvészeti szakértők tudásának ötvözésével nemcsak az általános jellegzetességek, hanem a tréningadatok feldolgozásával kapott speiális nyelvészeti jegyek is kezelhetők, ezáltal a programok hatékonysága nagyobb lehet. Növelve a tréningadatok méretét a program pontossága nőhet, cserélve a tréning adatokat a modulok speciális területekre hangolhatók.

3.1. A beolvasott szöveg szegmentálását végző modul

A beolvasott szöveg XML adatbázissá alakítása és az alapvető metainformációk (fejezet-, bekezdés-, mondat-, szóstruktúra) meghatározása a feldolgozás első lépcsője[2] [11]. A természetes nyelvi szövegekben számos különböző fajta szó jellegű, de a szótárakban nem szereplő lexikai elem található (szám, dátum, gépkocsirendszám, e-mail cím, stb.), amelyek felismerésére valamint a mondathatárok megállapítására egy formális-nyelvi eszközöket alkalmazó modul készült. A modul a GNU Flex [7] reguláris automatagenerátor eszközt használja. Ebben reguláris kifejezések írják le az ún. tokeneket (2. ábra). A *flex* program a reguláris kifejezésekből C programot készít, amely a szegmentáló modul magját alkotja. A mondatokra és a szavakra bontás hatásfoka igen jó, a hibásan felismert tokenek aránya nem több, mint 0,5%.

3.2. A szófaji elemző és egyértelműsítő modul

A programcsomagban több, különböző elven működő szófaji egyértelműsítő (*part-of-speech tagger*) modul található. Alapvetően minden egyes modul gépi tanuló algoritmussal meghatározott egyértelműsítési szabályokkal dolgozik – a

[7] A GNU Flex honlapja: `http://www.gnu.org/software/flex`

```
/* ponttal tagolt számok, pl. 12.000 */
NUMDOT [0-9]{1,3}("."[0-9]{3})+
NUMDIGIT ([0-9]+",")?[0-9]+
```

2. ábra. Reguláris kifejezések a Flex definíciós fájljában

különbség a felhasznált tanuló algoritmusokban van. A szabályok a többértelmű szó környezetében – előtte vagy utána – található más szavak morfo-szintaktikai kódjai, szótövei alapján hoznak döntést. Az egyik modul E. Brill TBL (Transformation Based Learning)[4] módszerén alapul, a második egy HMM (Hidden Markov Model) alapú algoritmus, amely a TnT tanuló módszert használja[3], a harmadik pedig egy logikai döntési szabályokat tanuló algoritmus[10], amelyet a tanszékcsoportban fejlesztettek ki. Fontosnak tartjuk megjegyezni, hogy bár az egyes tanuló módszerek forráskódjai nem állnak rendelkezésünkre, azonban a megtanult szabályokat végrehajtani képes programokból mindhárom esetben van saját forráskódú verzió is.

A modulok pontossága nagyon jónak mondható: az összes szóra vetített pontosság (*per-word accuracy*) 95,79% és 97,83% között mozgott[9]. Az információkinyerési alkalmazásokba bármelyik modul beépíthető, a modulok egymással teljes mértékben kompatibilisek. [8]

3.3. A felszíni szintaktikai elemző modul

A felszíni elemző feladata a mondatban szereplő főnévi struktúrák azonosítása. Ez a modul nem alkalmas egyéb bonyolultabb nyelvi szerkezetek, például határozói, jelzői, vagy igei szerkezetek felismerésére. Az elemző modul szintaktikai szabályait a Hócza András által készített RGLearn[8] algoritmus állítja elő. Az RGLearn a Szeged Korpusz 2.0-ból válogatott tréningadatok alapján Chomsky-féle formális szintaktikai szabályokat tanul.

A tanuló algoritmus a tréningadatokból gyűjtött főnéviszerkezet-fákból indul ki. Hasonló fákat keres, majd megpróbálja azokat egyesíteni. További eleme az algoritmusnak az ismétlődő, azonos morfo-szintaktikai kóddal rendelkező szavak sorozatának észlelése és egy általános rekurzív szabállyal történő helyettesítése. Azokat a régi szabályokat, amelyeket az újonnan megtanult általános szabályok magukba foglalnak, törli, és a folyamat addig folytatódik, amíg csak van lehetséges általánosítás.

A főnévi szerkezeteket felismerő modul a teljes Szeged Korpusz 2.0-n egy véletlen algoritmussal választott tréning- és attól független tesztadatok esetén 75,72% pontosságot ért el 81,69% találati arány mellett. Az üzleti rövidhírek esetén a pontosság 79,86% volt 86,63% találati aránnyal[8].

[8] A Szeged Korpusz hivatalosan az MSD szófaji kódolást használja, azonban a kutatók dolgoznak a Humor-kódolású verzión is. A tanszékcsoportban rendelkezésre áll kísérleti jelleggel mindhárom POS-tagger Humor-kódokkal tanított verziója is.

4. Az információkinyerés modellje és technológiája

Információkinyerő rendszerek megvalósításakor igyekeznek elkerülni a bonyolult szintaktikai szerkezetek azonosítását, mert így nagy mértékben gyorsítható a program működése. Az NKFP projekt résztvevői a feladatot mintaillesztési feladattá alakították át az alábbiakban ismertetésre kerülő módon.

4.1. Szemantikus keretek és felismerésük

A mondatok és a bennük leírt esemény absztrakt leírására bevezették a szemantikus keret (*semantic frame*) fogalmát. A szemantikus keret egy absztrakt eseménynek tekinthető, amely az írott szövegekben számos különböző megfogalmazásban (mondatban) fordulhat elő. A szemantikus keret tartalmaz egy fő cselekvést és ahhoz kapcsolódóan különböző szereplőket. A szereplők azonosítása lexikális, morfo-szintaktikai, felszíni elemzési és ontológiai attribútumaikra vonatkozó feltételekkel történik. Egy mondat akkor illeszthető egy szemantikus keretre, ha a keretben definiált fő ige és a szereplők a mondatbeliekkel mind egyenként azonosíthatók. A modul az előre kidolgozott szemantikus kereteket próbálja végig minden egyes mondaton. A szemantikus keretek további építőeleme az információs ablak (*information slot*). Amennyiben a keret illesztése sikeres, úgy az ablakban jelenik meg a keresett információ, amelyet a felhasználó keres.

A kezdeti kutatásokban a szemantikus keretek egyetlen mondatnak voltak megfeleltethetők, az információs ablakok pedig a szereplőkkel voltak azonosak. A rendszer továbbfejlesztése folyamatban van elsősorban a több mondattal megfogalmazott események egyetlen kerettel történő leírásának irányában[7].

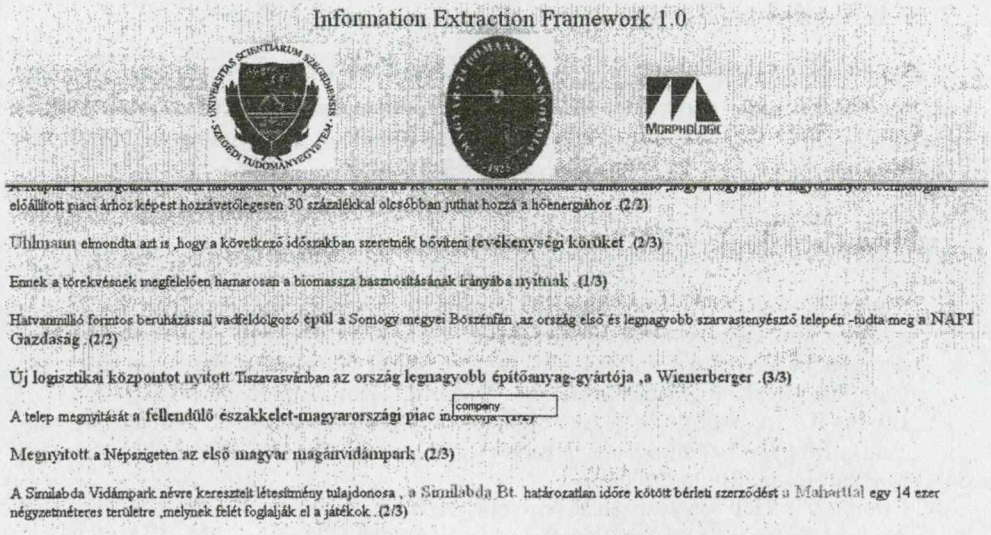
5. A felhasználói felület és a webes megjelenítő modul

A programcsomagban az utolsó modul a felhasználói felület, amely egy HTML nyelvű weblapot készít. Ez egyrészt tartalmazza a beolvasott nyers szöveget, másrészt a programcsomag által hozzáadott metainformációkat. Ez utóbbiakat grafikus eszközökkel jeleníti meg. A modatokban azonosított szereplők különböző színekkel, a szerepek nevei a weblapon lebegő üzenetablakokban (*tooltip*ekben) jelennek meg. A mondatokra illesztett szemantikus keret összes szereplőjének száma és az azonosított szereplők száma a mondatok után található. A 3. ábrán a webes megjelenítő modul egy képernyője látható.

Egy alternatív megjelenítő modul az eredményeket nem weblapon, hanem Excel-ablakban jeleníti meg. Az azonos eseményeket egy munkalapon, a mondatokat egy-egy sorban, az azonos szereplőket pedig azonos oszlopokban jeleníti meg. Ez a táblázat további feldolgozások kiindulópontja lehet.

6. Eredmények

Az NKFP projekt keretében ugyanannak az adatbázisnak és háttértudásnak a felhasználásával a konzorciumon belül két információkinyerésre alkalmas prog-



3. ábra. A webes megjelenítő modul egy képernyője

ramrendszer is készült. Az egyik [12] [13] nem különálló modulokból álló, pragmatikus, célratörő rendszer, amely azonban a folyamatos kísérletezést, a modulok cserélgetését nem támogatja. Az itt bemutatott rendszer e hiányosságokat igyekszik kiküszöbölni, és kifejezetten a további információkinyerési kutatások támogatására készült. Az Informatikai Tanszékcsoport Európai K+F pályázatokkal próbál támogatást szerezni a további kutatómunkához.

A cikkben bemutatott programcsomag tesztelésére a kutatók egy keretrendszert (benchmark) készítettek. Ez kézzel előre annotált, a rendszer számára ismeretlen mondatokat tartalmaz két előre kiválasztott témakörben: a tulajdonosváltás és az új telephely nyitása témakörében. [9] Erre a két témakörre megfelelő számú szemantikuskeret-definíció és rövidhír állt rendelkezésre. A tesztmondatok szöveges alakjára lefuttatták a programcsomag egyes komponenseit, majd összehasonlították a kézzel készített és a gép által előállított két állomány metainformációit. Tekintve, hogy a programcsomag több elemből áll, az egyes modulok hibája kumulálódik a végeredményben. Az eredmények megbízhatóbb értékelése érdekében arra is van lehetőség, hogy az egyes modulok által adott részeredményeket külön értékeljék, és összehasonlítsák az etalonfájllal.

A programcsomag által elért eredmény 70,2% (a pontosságból és a találati arányból számított kombinált érték). [10] A hibák 44%-át a felszíni elemző 29%-át a mondatmintákat felismerő program követte el. Tekintve, hogy a felszíni elemző egyelőre viszonylag sok hibát követ el, ha a hibás főnévi szerkezetek miatt bekövetkező mintallesztési hibákat nem számítjuk, akkor a pontosság jobb: 83,4% lesz.

[9] A benchmarkban 176 rövidhír, illetve 285 mondat szerepel.

[10] $F = \frac{2rp}{(r+p)}$, ahol r a találati arány, p a pontosság.

7. Köszönetnyilvánítás

A szerzők ezúton fejezik ki köszönetüket az OM NKFP 2/017/2001 projektbeli konzorciumi partnereiknek, az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztályának és a MorphoLogic Kft.-nek, akikkel a tudományos és szakmai kapcsolatokon túl szoros, személyes kapcsolatot alakítottak ki.

Hivatkozások

1. Alexin Z., Csirik, J., Gyimóthy, T., Bibok K., Hatvani, Cs., Prószéky, G., Tihanyi, L.: Manually Annotated Hungarian Corpus. in Proc. of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03, Budapest, Hungary, 53–56 (2003).
2. Bibok K.: A szóról és a szófajokról (a számítógépes nyelvfeldolgozás kapcsán), Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 31–36, (2003).
3. Brants, T.: TnT – a statistical part-of-speech tagger, in Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, USA, WA (2000).
4. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, **21** 543–565 (1995).
5. Csendes D., Csirik J., Gyimóthy T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus in Proc. of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), Brno, Czech Republic, 41–47, (2004).
6. Csendes, D., Hatvani, Cs., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz, Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 238–245, (2003).
7. Farkas R., Konczer K., Szarvas Gy.: Szemantikus keretillesztés és az IE rendszer automatikus kiértékelése Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004), beküldve, Szeged, Magyarország, (2004).
8. Hócza, A.: Noun Phrase Recognition with Tree Patterns elfogadva az Acta Cybernetica c. lapban történő megjelenésre (2004).
9. Kuba A., Hócza A., Csirik J.: POS Tagging of Hungarian with Combined Statistical and Rule-Based methods in Proc. of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), Brno, Czech Republic, 113–120, (2004).
10. Kuba A., Bakota T., Hócza A., Oravecz Cs.: A magyar nyelv néhány szófaji elemzőjének összevetése Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 16–22, (2003).
11. Miháczi András, Németh László, Rácz Miklós: Magyar szövegek temészetes nyelvi előfeldolgozása Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 38–43, (2003).
12. Prószéky G.: Automatikus információszerzés gazdasági rövidhírekből. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 161–166, (2003).
13. Prószéky G.: Automatikus információszerzés gazdasági-politikai rövidhírekből. VIII. Országos (Centenáriumi) Neumann Kongresszus kiadványa, Budapest, Magyarország, 359–367, (2003).